LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

UCRL-JRNL-206211

Distributions of pI vs pH provide prior information for the design of crystallization screening experiments

K. A. Kantardjieff, M. Jamshidian, B Rupp

August 24, 2004

Bioinformatics

Distributions of pI vs pH provide prior information for the design of crystallization screening experiments

Katherine A. Kantardjieff [1], Mortaza Jamshidian[2] and Bernhard Rupp[3,4]

[1] *Department of Chemistry and Biochemistry and W.M. Keck Foundation Center for Molecular Structure, California State University Fullerton, Fullerton, CA 92834-6866, USA*

[2] *Department of Mathematics, California State University Fullerton, Fullerton, CA 92834-6866, USA*

[3] *Department of Biochemistry and Biophysics, Texas A&M University, College Station, TX 77843-2128, USA*

[4] *Biology and Biotechnology Research Program, L-448, Lawrence Livermore National Laboratory Livermore, CA 94551, USA*

Correspondence: Phone: (714) 278-3752, Fax: (734) 939-4225
E-mail: kkantardjieff@fullerton.edu

Running title: pI as a crystallization predictor

Keywords: pI, isoelectric point, protein crystals, crystallization pH, high-throughput crystallography

In a recently published advanced access article, we described the use of a relationship between the distribution of pH of crystallization and the pI of the protein being crystallized to design more efficient crystallization screening experiments. The validity of our analysis has been questioned by Huber and Kobe. We wish to address their concern regarding our interpretation and clarify the electronic version of our article.

The motivation for our analysis is that although a number of studies have attempted to provide improved crystallization strategies (Gilliland et al. 2002), much of the 'knowledge' disseminated about protein crystallization continues to be anecdotal, with little statistical evidence or control experiments to prove general efficiency or usefulness. Considering the large number of physical, chemical and method related parameters, very few are sampled (and reported) with sufficient overlap to allow their direct use as a predictor for optimizing crystallization success (Rupp 2003). One parameter that is however frequently reported - regardless of the crystallization strategy employed - is the pH of the crystallization cocktail. Use of pH as a predictor for crystallization success, either globally or in correlation with the minimum solubility of a given protein at its isoelectric point, pI, appears attractive. Unfortunately, no direct correlation between minimum solubility at the pI and the pH of crystallization has ever been established (Samsudzi and Fivash 1992; Farr et al. 1998; Beretta et al. 2000). However, a significant relationship in fact exists between the calculated isoelectric point, pI, of successfully crystallized proteins with pI less than 7, and the reported pH at which they were most frequently crystallized, as we describe shortly.

Because pH is one of the few consistently reported parameters in the Protein Data Bank (PDB), we used the 9596 SEQRES records of a nonredundant protein data set (excluding

proteins/nucleic acid complexes) (Kantardjieff and Rupp 2003), which contain the entire expressed construct sequence including any tags, fusions or linkers, to calculate the pI using the pK_a values of Bjellqvist et al. (Bjellqvist and al 1993). The frequency distribution for pI of proteins is biomodal (Figure 1A), with highest frequencies (modes) at approximately pH 5.7 and 9.0, corresponding to the pI distribution for proteins encoded by sequenced genomes. (See for example (Urquhart et al. 1998; Baisnee et al. 2001; Adams et al. 2003).) The frequency distribution for reported crystallization pH of proteins is unimodal, with mean = 6.7, median = 6.9 and mode = 7.5 (Figure 1B). A similar crystallization pH distribution has been observed from unbiased random screening experiments in structural genomics initiatives (Hosfield et al. 2003; Rupp 2003; Rupp and Wang 2004).

In our original article, we reported that empirical distributions of the observed data imply a preferred range of crystallization pH for acidic and basic proteins, and these preferences provide strong prior information for the design of crystallization screening experiments of significantly increased efficiency. We thank Huber and Kobe for correctly pointing out that the correlation line depicted in Figure 2 of our original article is a result of our transformation of the data. Indeed, this confusing presentation has already been brought to our attention (Stewart 2003). However, the correlation reported erroneously in that figure was *never* used to produce a predictive model via regression analysis, which Huber and Kobe overlooked.

Transformation of the data preserves the experimental connection between a pI and its corresponding crystallization pH. In practice, the distributions of (pI vs pH) and (pI vs pI-pH) are identical, as verified by quantile-quantile plots. Huber and Kobe's distributions are not equivalent to ours, because the empirically observed pH and pI data (Figure 1) are clearly *not* uniformly distributed and *do* contain information. In fact, one observes statistically significant

non-zero correlations between pH and pI. Clustering of the bivariate observed data pI and pH, using the posterior probabilities from a mixture of Gaussian distributions, led us to assess the data in two separate groups, pI less than 7 (6266 cases) and pI greater than or equal to 7 (3330 cases). For each group, and for all the data together, we computed 99% confidence intervals for the Pearson correlation coefficient between pI and pH. These were (0.04, 0.10), (-0.08, 0.006), and (-0.003, 0.0496)[1] respectively, indicating that the correlation between pI and pH for the first group is significantly different from zero.

This statistical significance of the correlation allows a linear regression of pH on pI for the first group. SPSS was used to fit the linear regression model leading to $pH = 5.880 + 0.142 \times pI$. The standard errors of the intercept and the slope are 0.137 and 0.023 respectively, with both estimates being significantly different from zero (p-values <0 .0001). Based on the statistics related to this fit, a $100(1-a)\%$ confidence interval for the expected pH levels can be obtained using

$$5.88 + (0.142 \times pI) \pm z_{a/2} \sqrt{1.536/6266 + (pI - 5.82)^2 \times (.023)^2},$$

where $z_{a/2}$ is the $1-a/2$ quantile of the standard normal distribution. Common values of $z_{a/2}$ for 90%, 95%, and 99% confidence are 1.64, 1.96, and 2.57 respectively. This, for example, predicts that in repeated experiments with $pI = 5$ the mean of the expected crystallization pH values will be between 6.54 and 6.64 with 95% confidence. The corresponding

[1] The confidence intervals were computed based on the distribution of the Fisher transformation of the Pearson correlation. We also computed these intervals using the bootstrap method, and the resulting intervals were in agreement.

prediction interval, indicating the predicted pH value for a given pI in a single experiment, can be obtained using

$$5.88 + (0.142 \times pI) \pm z_{a/2} \sqrt{1.536 + 1.536/6266 + (pI - 5.82)^2 \times (.023)^2}.$$

For $pI = 5$ the 95% prediction interval is (4.2, 9.0), indicating that in approximately 95% of the successful experiments the pH value will be in the range 4.2 to 9.0. The 50% prediction interval for $pI = 5$ is narrower (5.8 to 7.4). Based on the fact that the correlation between pH and pI is not significant at the 1% level for cases other than the acidic proteins, we did not implement a predictive correlation model, but decided to follow the corresponding empirical distributions of pI vs pH (equivalent to pI vs pI-pH) shown in Figure 2.

CrysPred (http://www-structure.llnl.gov/cryspred/), the simple prototype pH range calculator described in our original article, shows how empirical distributions of observed crystallization pH for a given pI range can be used as prior information to optimize efficiency of initial crystallization screening in HTPX by identifying with the highest *overall* efficiency (least material, supplies and resources, and thus cost) the proteins that are most likely to yield useful or suitable crystals and structures. Results can be easily imported into any customizable screen generator that allows to define the frequency of occurrence for selected pH ranges (for example, CrysTool (Segelke and Rupp 1998; Segelke 2001). Supplemental information on the web site discusses additional caveats regarding accuracy of pI calculations and reported pH data, and usage bias of the experimental frequency distributions (Rupp and Wang 2004).

Crystallization is a special case of phase separation from a thermodynamically metastable solution under kinetic control (Rupp 2003). As discussed in our original article, while control over kinetic parameters such as nucleation or growth rates is rather difficult to achieve, attractive interaction between molecules as a thermodynamically necessary – but not sufficient – condition

for crystallization can be described on the basis of thermodynamic excess properties, such as their manifestation in the second virial coefficient, B_{22}. pH and electrolytes have been shown to be crucial solvent parameters that modulate potentials through specific and non-specific effects (Benas et al. 2002; Retailleau et al. 2002). Because charge distributions on proteins are discrete, and distances between charged residues on a protein surface are not negligible compared to the protein diameter, the specificity of charge interactions cannot be ignored, and colloidal model systems (Belloni 2000; Frenkel 2002) have not been successful in describing these interactions in proteins (Piazza 2000; Tardieu et al. 2002; Allahyarov et al. 2003).

While methods for choosing protein crystallization conditions have largely been empirical (McPherson 1982), with the recent development of random screening methods (Segelke and Rupp 1998; Segelke 2001), it is anticipated that statistical analysis will provide predictive frameworks that increase the probability of producing high quality crystals (Rupp and Wang 2004). Empirical distributions of observed PDB data for acidic and basic proteins, as well as for smaller groups of data binned by pI, suggest that acidic proteins crystallize with highest likelihood ~0-2.5 pH units *above* their isoelectric point, whereas basic proteins preferably crystallize ~0.5-3 pH units *below* their isoelectric point (with the sum of these related distributions adding up to the distribution shown in Figure 1). This conclusion is based on experimental facts (the pI vs pH distribution of observed data), and Huber and Kobe's statement "We have demonstrated here that this conclusion is based on misinterpretation of data and should not be used to guide crystallization experiments until a correlation between pH and pI is established" is untenable. Following an experimental distribution suitable for the selected pI range to increase overall efficiency is completely valid. Furthermore, for the pI less than 7 group, a correlation with crystallization pH has been established, which could be used to support a

predictive model. The absence of negatives in the PDB data does not allow calculation of actual propensities, however, and it is possible that the experimentally observed distributions are biased by usage (Rupp and Wang 2004). Thus, a predictive model based on correlations would be no more reliable at this point than following the empirical distributions of pH for a given range of pI. Our statistical analysis suggests favorable pH distributions and ranges for improved efficiency of crystallization screening experiments, validating early grid designs that varied pH around physiological values (McPherson 1982).

LLNL is operated by University of California for the US DOE under contract W-7405-ENG-48. This work was funded by NIH P50 GM62410 (TB Structural Genomics) center grant.

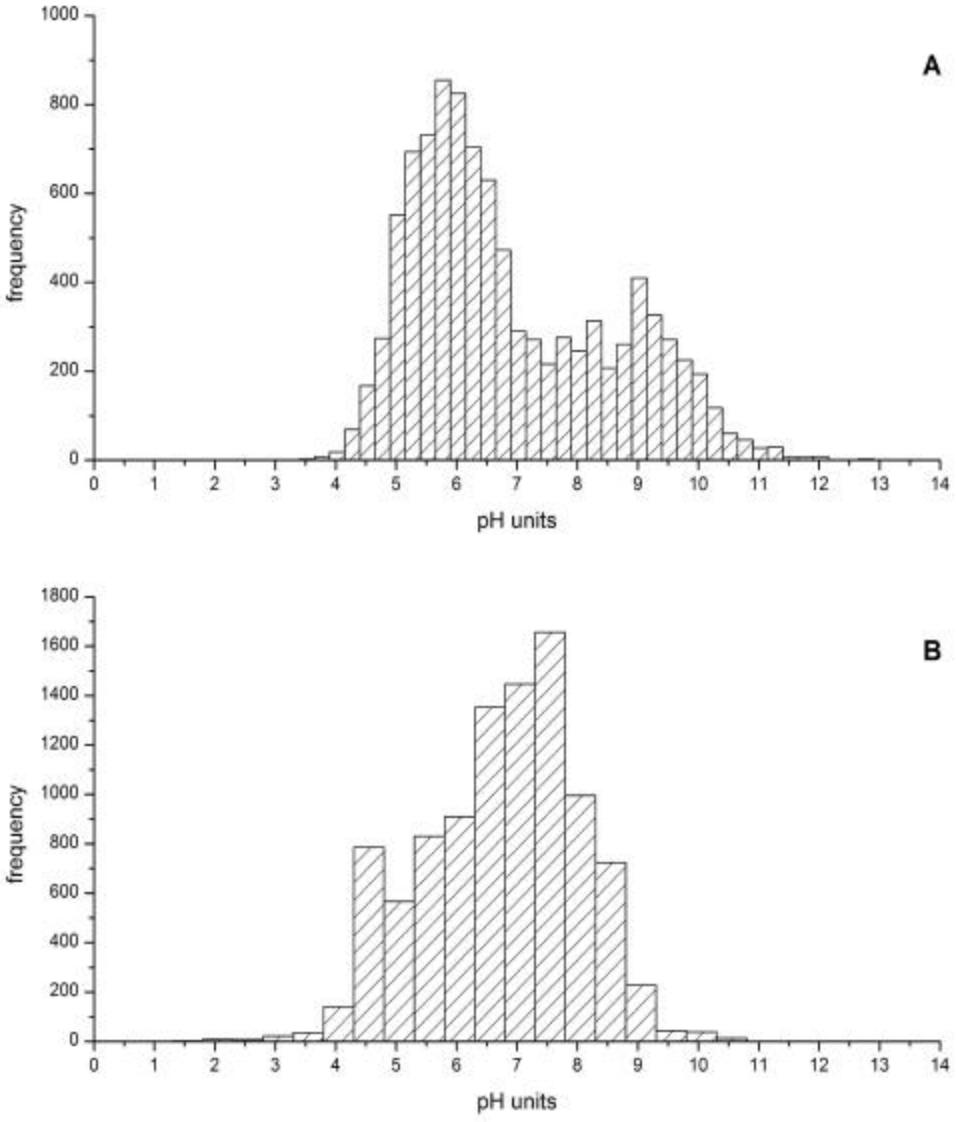

Figure 1 Frequency distributions for empirically observed PDB pH and pI data. (A) pI of successfully crystallized proteins. (B) reported pH of crystallization for proteins.

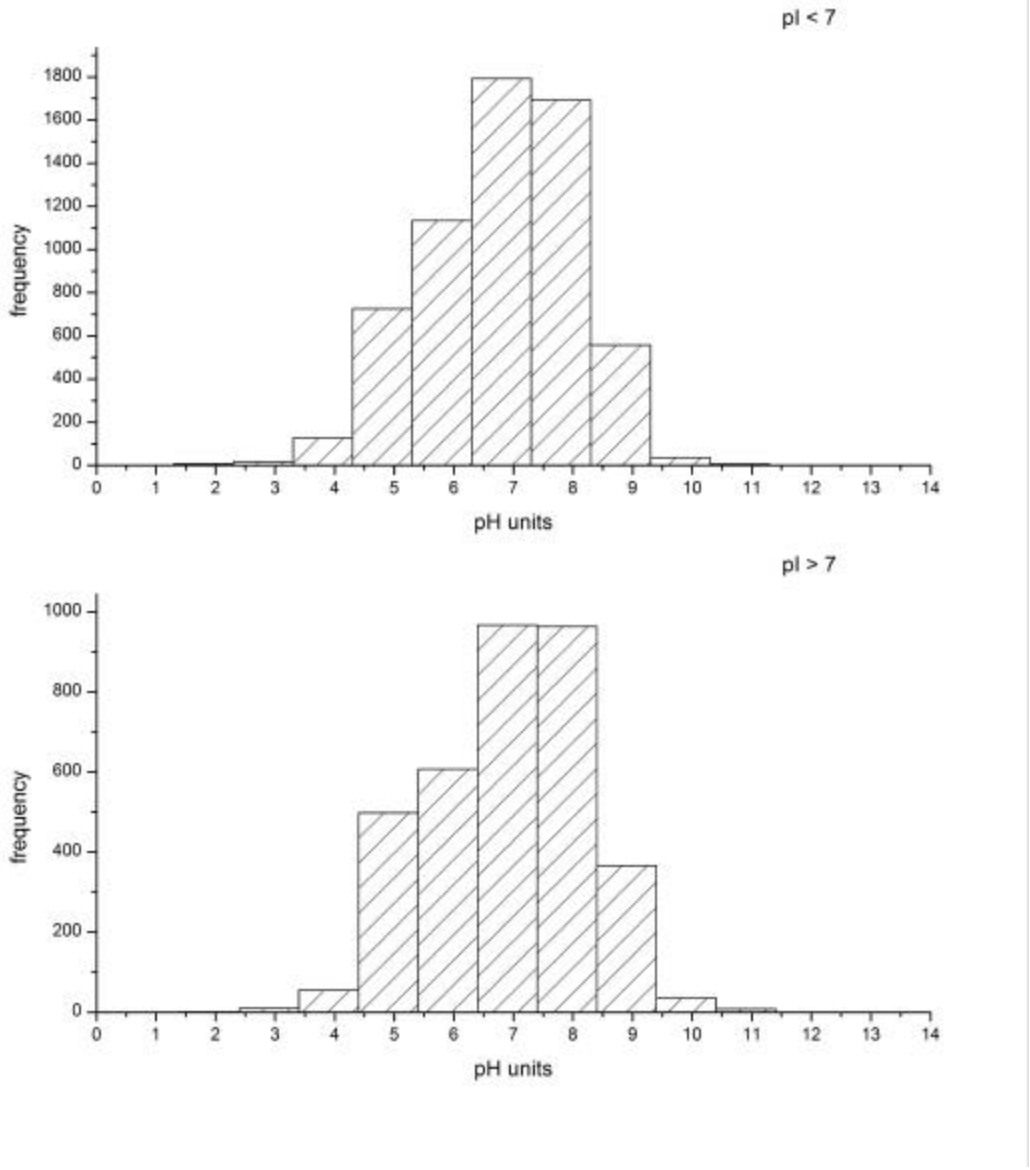

Figure 2 Frequency distribution of pH versus pI for successfully crystallized proteins. Top panel shows the frequency distribution of crystallization pH and calculated pI of successfully crystallized acidic proteins. Bottom panel shows this frequency distribution for basic proteins. Based on these distributions of *empirically observed data,* acidic proteins tend to crystallize 0-2.5 pH units *above* their pI, whereas basic proteins prefer to crystallize 0.5-3 pH units *below* their pI.

References

Adams, M.W., Dailey, H.A., DeLucas, L.J., Luo, M., Prestegard, J.H., Rose, J.P., and Wang, B.C. 2003. The Southeast Collaboratory for Structural Genomics: a high-throughput gene to structure factory. *Accounts Chem Res.* **36:** 191-198.

Allahyarov, E., Loewen, H., Hansen, J.P., and Louis, A.A. 2003. Nonmonotonic variation with salt concentration of the second virial coefficient in protein solutions. *Phys Rev E* **67:** 0514041-05140413.

Baisnee, P.-F., Baldi, P., Brunak, S., and Pedersen, A.G. 2001. Flexibility of the genetic code with respect to DNA structure. *Bioinformatics* **17:** 237-248.

Belloni, L. 2000. Colloidal interactions. *J Phys: Condens Matter* **12:** R549-R587.

Benas, P., Legrand, L., and Ries-Kautt, M. 2002. Strong and specific effects of cations on lysozyme chloride solubility. *Acta Cryst* **D:** 1582-1587.

Beretta, S., Chirico, G., and Baldini, G. 2000. Short-tange interactions of globular proteins at high ionic strengths. *Macromol* **33:** 8663-8670.

Bjellqvist, B., and al, e. 1993. The focusing positions of polypeptides in immobilized gradients can be predicted from their amino acid sequences. *Electrophoresis* **14:** 1023-1031.

Farr, R.G., Perryman, A.L., and Samsudzi, C.T. 1998. Re-clustering the database for crystallization of macromolecules. *J Cryst Growth* **183:** 653-658.

Frenkel, D. 2002. Soft condensed matter. *Physica A* **313:** 1-31.

Gilliland, G.L., Tung, M., and Ladner, J.E. 2002. The Biological Macromolecule Crystallization Database: crystallization procedures and strategies. *Acta Crystallogr D* **58:** 916-920.

Hosfield, D., Palan, J., Hilgers, M., Scheibe, D., McRee, D.E., and Stevens, R.C. 2003. A fully integrated protein crystallization platform for small-molecule drug discovery. *J Struct Biol* **142:** 207-217.

Kantardjieff, K.A., and Rupp, B. 2003. Matthews coefficient probabilities: Improved estimates for unit cell contents of proteins, DNA, and protein-nucleic acid complex crystals. *Prot Sci* **12**.

McPherson, A. 1982. *Preparation and analysis of protein crystals*. Wiley, New York.

Piazza, R. 2000. Interactions and phase transitions in protein solutions. *Curr Opin Colloid Interface Sci* **5:** 38-43.

Retailleau, P., Ducruix, A., and Ries-Kautt, M. 2002. Importance of the nature of anions in lysozyme crystallization correlated with protein net charge variation. *Acta Crystallogr D* **58:** 1576-1581.

Rupp, B. 2003. Maximum-likelihood crystallization. *J Struct Biol* **142:** 162-169.

Rupp, B., and Wang, J.-W. 2004. Predictive models for protein crystallization. *Methods***:** in press.

Samsudzi, C.T., and Fivash, M.J. 1992. Cluster analysis of the Biological Macromolecule Crystallization Database. *J Cryst Growth* **123:** 47-58.

Segelke, B., and Rupp, B. 1998. Beyond the Sparse Matrix Screen: A Web Service for Randomly Generating Crystallization Experiments. *ACA Meeting Series* **25:** 78.

Segelke, B.W. 2001. Efficiency analysis of sampling protocols used in protein crystallization screening. *J Crystal Growth* **232:** 553-562.

Stewart, P.S. 2003. Personal communication.

Tardieu, A., Bonnete, F., Finet, S., and Vivares, D. 2002. Understanding salt or PEG induced attractive interactions to crystallize biological macromolecules. *Acta Crystallogr D* **58:** 1549-1553.

Urquhart, B.L., Cordwell, S.J., and Humphrey-Smith, I. 1998. Comparison of predicted and observed properties of proteins encoded by the genome of Mycobacterium tuberculosis H37Rv. *Biochem Biophys Res Comm* **253:** 70-79.